Lab 1: Question 2

Connor McCormick, Max Hoff, Chi Ma

6/22/2021

Contents

1 Who Had a Harder Time Voting in the 2020 Election?

1.1 Importance and Context

Did Democratic voters or Republican voters report experiencing more difficulty voting in the 2020 election?

The 2020 election was arguably one of the most consequential elections in history. With the previous 4 years of the presidency being held by former president Donald Trump, a tremendous amount of change has occurred in our country that no doubt contributed to the significance of the election. One important shift we can observe is the advancement of technology, especially in the social media space. As social media algorithms have become more and more fine tuned, they were able to curate remarkably specific content for their users, leading to siloed content distribution that has been proposed to have caused a polarizing shift in political ideologies across the country. Not to mention, the growing adoption of technology nationwide led to a fake news epidemic where it became notably difficult to tell if what you were reading were true. Additionally, the introduction of the global pandemic 8 months before the election added a whole other layer of complexity. All of these happenings led to the past election becoming vastly different than any election prior. Although it may have been measurable in past elections, the question of which political party had more difficulty voting in the past election is more relevant than ever, due to the nature of events that unfolded.

1.2 Description of Data

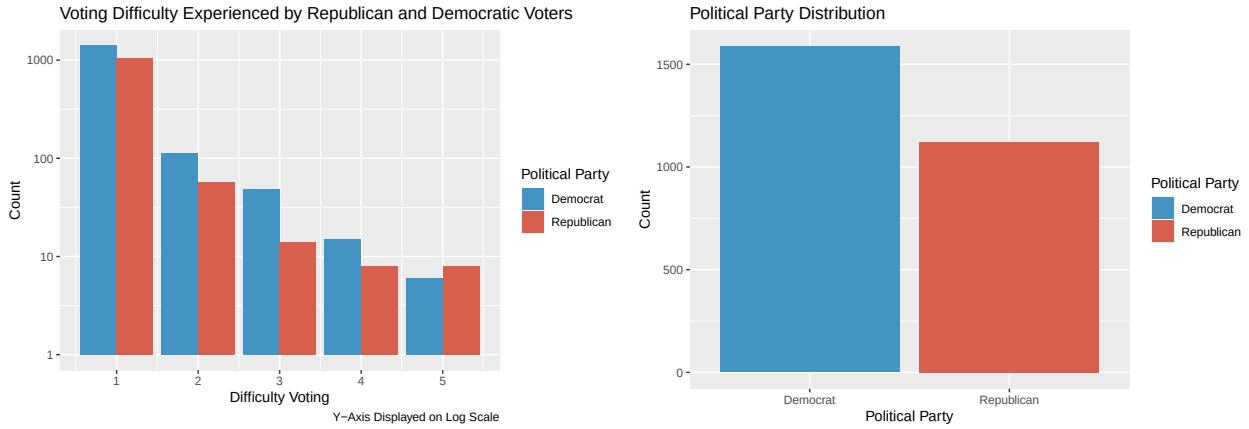
We will approach this question using data from the 2020 American National Election Study (ANES). This is an observational data set based on sample of survey respondents from the YouGov platform.

To start, I would like to give an overview of the process taken to acquire the sample that we will perform the hypothesis test with. We first selected four columns from the total data available to us: '2020 Case ID', 'Voter Turnout in 2020', 'Party of Registration' (Pre-election), and 'How Difficult was it to Vote'. We then labeled these columns with corresponding related variable names. Starting from the bottom, we selected the voting difficulty column since it is the response variable in the given research question. Similarly, we selected the party registration column because it is the explanatory variable. We selected the voter turnout column in order to subset the data by respondents who voted only. We noted that there likely would not have been a relevant entry in the voting difficulty column if the respondent hadn't voted, however we felt that it was a simple enough double check to put in place. Lastly, we selected the 2020 Case ID column in order to keep a unique identifier for respondents. Though it is unnecessary, it is a best practice and ended up in our sample as a force of habit.

Next, we will address the sample that we distilled and will perform our hypothesis test with. In order to reach the sample we have come to as shown in the Political Party Affiliation plot below, we first subset by respondents who reported a voting difficulty > 0. This was because any negative responses were reasons why the respondent did not give a valid response. Next, we subset by party affiliation, including respondents registered to either the Republican or Democrat party. It is important to note that although there was other data available that gave information on voter party affiliation (such as voter lean or how strongly they identify with a political party), we decided to perform our test using solely party registration data as this is a very unambiguous divide between the populations we are comparing and our sample sizes for both parties is over 1000, thus allowing the CLT to kick in. As you can see in the plots below, there is a relatively similar number of survey respondents registered to both the Democratic and Republican parties though there is a slightly larger number survey respondents registered Democrat. You will also notice that there is quite a large number of survey respondents shown in the 'Unknown' category for party affiliation (Shown in Table 1). These respondents either registered independent, registered other, or did not respond for another reason, so we did not include them in the sample.

Table 1: Party Affiliations in Sample

	Democrat	Republican	Unknown
Count	1587	1122	3696



Now, regarding the sample we will use to perform the hypothesis test, it is important to mention that the explanatory variable, political party affiliation, is a binary variable taking on values of either 'Republican' or 'Democrat'. The response variable, voting difficulty, is an ordinal variable structured on a Likert Scale.

Finally, we consider the voting difficulty responses. As shown in tables above, the majority response to the survey question: "How difficult was it for [respondent] to vote?", is 1 - "Not difficult at all". As the response variable gets larger, the number of respondents decreases, with Democrats consistently taking up a larger proportion of each response until the final response: 5 - "Extremely difficult". This leads us to hypothesize that there will be a difference in the voting difficulty experienced between parties. However, we must conduct the hypothesis test to find out if there is statistically significant evidence of this effect.

1.3 Hypothesis

Null Hypothesis: There was no difference in the reported voting difficulty experienced between Democrats and Republicans in the 2020 election. Stated in statistical terms, if we let X = the amount of voting difficulty experienced by Democrats and we let Y = the amount of voting difficulty experienced by Republicans, then the null hypothesis would be that P(X > Y) = P(X < Y).

Alternative Hypothesis: There was a difference in the reported voting difficulty experienced between Democrats and Republicans in the 2020 election. Again stated in statistical terms, if we follow the same variable assignments of X and Y laid out in the null hypothesis, we would have the alternative hypothesis be that P(X > Y) != P(X < Y)

1.4 Most appropriate test

When deciding which statistical test is the most appropriate in this case, we first identify the types of variables we are working with. As mentioned before, the political party variable is a binary variable taking on either a 'Democrat' or a 'Republican' value. On the other hand, the difficulty voting variable is an ordinal variable from 1 to 5. At this point we decide that we need a non-parametric test. Next we note that this data is unpaired (an easy way to tell is that our two groups have different sizes), so we conclude that the most appropriate test to perform will be the Wilcoxon Rank-Sum Test.

Just to double check that our data is in line with the assumptions for the Wilcoxon Rank-Sum Test, we will go through each of them. The first assumption is that our data is i.i.d. The ANES 2020 data set collected

responses through three methods: web-only (online survey), mixed web (online survey + phone interview), and mixed video (online survey + video interview + phone interview). While there is a small chance of dependencies occurring with this design (i.e. one respondent encourages his friend to complete the survey as well), these occurrences are rare due to the size of the survey response rate (36.7% overall) and the population sampled (231 million voters over 18). The next assumption is that we are working with ordinal variables. As mentioned before, the response variable voting difficulty is in fact measured on the ordinal scale which allows us to compare the two distributions since we are separating on the explanatory variable political party. Since these two assumptions hold with our data sample, we are able to move forward and perform the Wilcoxon Rank-Sum Test

1.5 Test, results and interpretation

```
Democrats <- voting_diff %>%
  subset(political_party == 'Democrat') %>%
  select(difficulty_voting)

Republicans <- voting_diff %>%
  subset(political_party == 'Republican') %>%
  select(difficulty_voting)

wilcox_test <- wilcox.test(Democrats$difficulty_voting, Republicans$difficulty_voting,
            data = voting_diff,
            paired=FALSE,
            exact = FALSE)

correlation <- cor.test(as.numeric(voting_diff_sub$difficulty_voting),
                        as.numeric(voting_diff_sub$political_party),
                        method='spearman')

pairings <- sum(voting_diff_sub$political_party &
                      !is.na(voting_diff_sub$difficulty_voting), na.rm=T) *
            sum(! voting_diff_sub$political_party  &
                      !is.na(voting_diff_sub$difficulty_voting) , na.rm=T)
```

Based on the result of the test, there is enough evidence with a p-value of 0.0017 to reject the null-hypothesis. Thus, we can conclude that the probability that Democrats experienced more voting difficulty than Republicans is not equal to the probability that Republicans experienced more voting difficulty than Democrats. In order to express this result practically we will calculate the rank biserial correlation. This correlation uses rank information in the difficulty voting variable to measure how much political party varies with difficulty voting. In this case, we observe that the correlation coefficient is only -0.06, which suggests that the effect size is rather small. Additionally, we note that there is a test_statistic of W = 922827, which we will use to calculate the proportion of pairings where one party had more difficulty than another voting. Since the Democrat data vector was our x variable and the Republican our y in the Wilcoxon Rank Sum test, we have 922827/1780614 = 0.518, so we can say that in 51.8% of pairings between a Democratic voter and a Republican voter, the Democratic voter will report more difficulty voting. So even though the effect size is rather small, we can still conclude with significant evidence that Democratic voters had a slightly more difficult time voting in the 2020 election.